

What is claimed is:

1 1. A computer-implemented method of determining discourse
2 structures, the method comprising:
3 generating a set of one or more discourse parsing decision
4 rules based on a training set; and
5 determining a discourse structure for an input text segment
6 by applying the generated set of discourse parsing decision rules
7 to the input text segment.

1 2. The method of claim 1 wherein the training set comprises
2 a plurality of annotated text segments and a plurality of
3 elementary discourse units (EDUs), each annotated text segment
4 being associated with a set of EDUs that collectively represent
5 the annotated text segment.

1 3. The method of claim 2 wherein the annotated text
2 segments are built manually by human annotators.

1 4. The method of claim 2 wherein generating the set of
2 discourse parsing decision rules comprises iteratively performing
3 one or more operations on a set of EDUs to incrementally build
4 the annotated text segment associated with the set of EDUs.

1 5. The method of claim 4 wherein the one or more operations
2 iteratively perform comprise a shift operation and/or one or more
3 reduce operations.

1 6. The method of claim 5 wherein the reduce operations
2 comprise one or more of the following six operations: reduce-ns,
3 reduce-sn, reduce-nn, reduce-below-ns, reduce-below-sn, reduce-
4 below-nn.

1 7. The method of claim 5 wherein the six reduce operations
2 and the shift operation are sufficient to derive the discourse
3 tree of any input text segment.

1 8. The method of claim 1 wherein determining a discourse
2 structure comprises incrementally building a discourse tree for
3 the input text segment.

1 9. The method of claim 8 wherein incrementally building a
2 discourse tree for the input text segment comprises selectively
3 combining elementary discourse trees (EDTs) into larger discourse
4 tree units.

1 10. The method of claim 8 wherein incrementally building a
2 discourse tree for the input text segment comprises performing
3 operations on a stack and an input list of elementary discourse

4 trees (EDTs), one EDT for each elementary discourse unit (EDU) in
5 a set of EDUs corresponding to the input text segment.

1 11. The method of claim 10 further comprising, prior to
2 determining the discourse structure for the input text segment,
3 segmenting the input text segment into EDUs and inserting the
4 EDUs into the input list.

1 12. The method of claim 1 wherein determining the discourse
2 structure for the input text segment further comprises:

3 segmenting the input text segment into elementary discourse
4 units (EDUs);

5 incrementally building a discourse tree for the input text
6 segment by performing operations on the EDUs to selectively
7 combine the EDUs into larger discourse tree units; and

8 repeating the incremental building of the discourse tree
9 until all of the EDUs have been combined.

1 13. The method of claim 12 wherein segmenting the input
2 text segment into EDUs is performed by applying a set of
3 automatically learned discourse segmenting decision rules to the
4 input text segment.

1 14. The method of claim 13 further comprising generating
2 the set of discourse segmenting decision rules by analyzing a
3 training set.

1 15. The method of claim 1 wherein the input text segment
2 comprises a clause, a sentence, a paragraph or a treatise.

1 16. A computer-implemented text parsing method comprising:
2 generating a set of one or more discourse segmenting
3 decision rules based on a training set; and
4 determining boundaries in an input text segment by applying
5 the generated set of discourse segmenting decision rules to the
6 input text segment.

1 17. The method of claim 16 wherein determining boundaries
2 comprises examining each lexeme in the input text segment in
3 order.

1 18. The method of claim 17 further comprising assigning,
2 for each lexeme, one of the following designations: sentence-
3 break, EDU-break, start-parenthetical, end-parenthetical, and
4 none.

1 19. The method of claim 17 wherein examining each lexeme in
2 the input text segment comprises associating features with the
3 lexeme based on surrounding context.

1 20. The method of claim 16 wherein determining boundaries
2 in the input text segment comprises recognizing sentence
3 boundaries, elementary discourse unit (EDU) boundaries,
4 parenthetical starts, and parenthetical ends.

1 21. A computer-implemented method of generating discourse
2 trees, the method comprising:
3 segmenting an input text segment into elementary discourse
4 units (EDUs); and
5 incrementally building a discourse tree for the input text
6 segment by performing operations on the EDUs to selectively
7 combine the EDUs into larger discourse tree units.

1 22. The method of claim 21 further comprising repeating the
2 incremental building of the discourse tree until all of the EDUs
3 have been combined into a single discourse tree.

1 23. The method of claim 21 wherein the incremental building
2 of the discourse tree is based on predetermined decision rules.

1 24. The method of claim 23 wherein the predetermined
2 decision rules comprise automatically learned decision rules.

1 25. The method of claim 23 further comprising generating
2 the predetermined decisions rules by analyzing a training set of
3 annotated discourse trees.

1 26. The method of claim 21 wherein the operations performed
2 on the EDUs comprise one or more of the following: shift, reduce-
3 ns, reduce-sn, reduce-nn, reduce-below-ns, reduce-below-sn,
4 reduce-below-nn.

1 27. A discourse parsing system comprising:
2 a plurality of automatically learned decision rules;
3 an input list comprising a plurality of elementary discourse
4 trees (EDTs), each EDT corresponding to an elementary discourse
5 unit (EDU) of an input text segment;
6 a stack for holding discourse tree segments while a
7 discourse tree for the input text segment is being built; and
8 a plurality of operators for incrementally building the
9 discourse tree for the input text segment by selectively
10 combining the EDTs into a discourse tree segment according to the
11 plurality of decision rules and moving the discourse tree segment
12 onto the stack.

1 28. The system of claim 27 further comprising a discourse
2 segmenter for partitioning the input text segment into EDUs and
3 inserting the EDUs into the input list.

1 29. A computer-implemented method comprising determining a
2 discourse structure for an input text segment by applying a set
3 of automatically learned discourse parsing decision rules to an
4 input text segment.

1 30. A computer-implemented summarization method comprising:
2 generating a set of one or more summarization decision rules
3 based on a training set; and
4 compressing a tree structure by applying the generated set
5 of summarization decision rules to the tree structure.

1 31. The method of claim 30 wherein the tree structure
2 comprises a discourse tree.

1 32. The method of claim 30 wherein the tree structure
2 comprises a syntactic tree.

1 33. The method of claim 30 further comprising generating
2 the tree structure to be compressed by parsing an input text
3 segment.

1 34. The method of claim 33 wherein the input text segment
2 comprises a clause, a sentence, a paragraph, or a treatise.

1 35. The method of claim 30 further comprising converting
2 the compressed tree structure into a summarized text segment.

1 36. The method of claim 35 wherein the summarized text
2 segment is grammatical and coherent.

1 37. The method of claim 35 wherein the summarized text
2 segment includes sentences not present in a text segment from
3 which the pre-compressed tree structure was generated.

1 38. The method of claim 30 wherein applying the generated
2 set of summarization decision rules comprises performing a
3 sequence of modification operations on the tree structure.

1 39. The method of claim 38 wherein the sequence of
2 modification operations comprises one or more of the following: a
3 shift operation, a reduce operation, and a drop operation.

1 40. The method of claim 39 wherein the reduce operation
2 combines a plurality of trees into a larger tree.

1 41. The method of claim 39 wherein the drop operation
2 deletes constituents from the tree structure.

1 42. The method of claim 30 wherein the training set
2 comprises pre-generated long/short tree pairs.

1 43. The method of claim 42 wherein generating the set of
2 summarization decision rules comprises iteratively performing one
3 or more tree modification operations on a long tree until the
4 paired short tree is realized.

1 44. The method of claim 43 wherein a plurality of
2 long/short tree pairs are processed to generate a plurality of
3 learning cases.

1 45. The method of claim 44 wherein generating the set of
2 decision rules comprises applying a learning algorithm to the
3 plurality of learning cases.

1 46. The method of claim 44 further comprising associating
2 one or more features with each of the learning cases to reflect
3 context.

1 47. A computer-implemented summarization method comprising:
2 generating a parse tree for an input text segment; and
3 iteratively reducing the generated parse tree by selectively
4 eliminating portions of the parse tree.

1 48. The method of claim 47 wherein the generated parse tree
2 comprises a discourse tree.

1 49. The method of claim 47 wherein the generated parse tree
2 comprises a syntactic tree.

1 50. The method of claim 47 wherein the iterative reduction
2 of the parse tree is performed based on a plurality of learned
3 decision rules.

1 51. The method of claim 47 wherein iteratively reducing the
2 parse tree comprises performing tree modification operations on
3 the parse tree.

1 52. The method of claim 51 wherein the tree modification
2 operations comprise one or more of the following: a shift
3 operation, a reduce operation, and a drop operation.

1 53. The method of claim 52 wherein the reduce operation
2 combines a plurality of trees into a larger tree.

1 54. The method of claim 52 wherein the drop operation
2 deletes constituents from the tree structure.

1 55. A computer-implemented summarization method comprising:
2 parsing an input text segment to generate a parse tree for
3 the input segment;
4 generating a plurality of potential solutions;
5 applying a statistical model to determine a probability of
6 correctness for each of potential solution;
7 extracting one or more high-probability solutions based on
8 the solutions' respective determined probabilities of
9 correctness.

1 56. The method of claim 55 wherein the generated parse tree
2 comprises a discourse tree.

1 57. The method of claim 55 wherein the generated parse tree
2 comprises a syntactic tree.

1 58. The method of claim 55 wherein applying a statistical
2 model comprises using a stochastic channel model algorithm.

1 59. The method of claim 58 wherein using a stochastic
2 channel model algorithm comprises performing minimal operations
3 on a small tree to create a larger tree.

60. The method of claim 58 wherein using a stochastic channel model algorithm comprises probabilistically choosing an expansion template.

61. The method of claim 55 wherein generating a plurality of potential solutions comprises identifying a forest of potential compressions for the parse tree.

62. The method of claim 61 wherein the generated parse tree has one or more nodes, each node having N children (wherein N is an integer), and wherein identifying a forest of potential compressions comprises:

generating $2^N - 1$ new nodes, one node for each non-empty subset of the children; and

packing the newly generated nodes into a whole.

63. The method of claim 61 wherein the generated parse tree has one or more nodes, and wherein identifying a forest of potential compressions comprises assigning an expansion-template probability to each node in the forest.

64. The method of claim 55 wherein extracting one or more high-probability solutions comprises selecting one or more trees based on a combination of each tree's word-bigram and expansion-template score.

1 65. The method of claim 64 wherein selecting one or more
2 trees comprises selecting a list of trees, one for each possible
3 compression length.

1 66. The method of claim 55 further comprising normalizing
2 each potential solution based on compression length.

1 67. The method of claim 55 further comprising, for each
2 potential solution, dividing a log-probability of correctness for
3 the solution by a length of compression for the solution.